메타 검색에서 질의와 컬렉션 사이의 관련성 분포정보를 이용한 컬렉션 선택

배종민[†] · 김현주[††]

요 약

메타 검색에서 이질의 컬렉션으로부터 정보를 검색할 때, 주어진 질의에 대하여 가장 적합한 컬렉션을 선택하는 것에 대한 알고리즘을 제안한다. 제안된 컬렉션 선택 방법은 검색에 참여한 컬렉션으로부터 질의에 대해 임의의 크기 N 만큼 검색 문서를 수집한 후에 이를 분석하여 컬렉션에 대한 관련성 정도로 추정하고 이를 기반으로 컬렉션 선택 기준을 결정하였다. 이때 가장 적합한 컬렉션을 선택하기 위해서, 모집단의 크기 N, 관련문서의 순서정보, 정확도 등의 메타 정보를 사용하였다.

Collection Selection using Relevance Distribution Information between Queries and Collections in Meta Search

Jong-Min Bae[†] and Hyun-Ju Kim[††]

ABSTRACT

This paper proposes an efficient algorithm to select the proper retrieval results from various information sources in Meta search. The algorithm collects and evaluates the related documents to the given query. Then, it determines the appropriate retrieval results based on the relevance between the query and the collected documents. This algorithm depends on the Meta information such as the size N of population, top-ranked information of related documents and the precision in order to choose the most appropriate retrieval result.

1. 서 론

최근 컴퓨터의 급속한 보급과 함께 컴퓨터 네트워크 환경이 매우 폭넓게 사용되고 있다. 그 중에서도 인터넷으로 이용할 수 있는 컬렉션의 종류는 우리들의 상상을 초월하고 있다. 이와 같은 인터넷상의 컬렉션들은 매우 다양하게 생겨났으며 지금도 개발되고 있다. 이러한 컬렉션들로부터 사용자가 원하는 정보를 얻는 방법으로는 컬렉션이 제공하는 검색 엔진을 사용하여 찾고자하는 정보를 얻는다. 그러나 이는 인터넷상에 존재하는 수많은 컬렉션 가운데서 자신이 원하는 정보가 어디에 있는지 찾기도 힘들뿐만 아니라 또한 찾았다 하더라도 컬렉션의 검색 엔진을 효과적으로 사용하기도 어렵다[14-17].

최근에는 이들 컬렉션이 가지고 있는 검색 엔진들을 사용자가 쉽고 편리하게 이용할 수 있도록 하는 정보검색 분야의 노력 중 하나가 통합 검색(Federated Search) 혹은 메타 검색(Meta Search)의 등장이다. 최근 등장한 메타 검색기로서는 ProFusion, SavvySearch, 미스다찾니 등이 있으며, 기존의 Yahoo, InfoSeek 등과 같은 정보 검색 시스템도 폭발적으로 늘어나는 정보를 자신의 컴퓨터에 저장하는 중앙 집중식 정보 관리법에 한계를 느껴 InfoSeek에서는 메타 검색 시스템인 InfoSeek Patent를 개발하여 실험적으로 운영하고 있다[9,10,11,13].

이러한 메타 검색 분야에서 질의에 대해 검색 결과에 영향을 미치는 주요 요인으로써는 세 가지 연구 분야로 분류할 수 있다. 첫 번째는 가장 좋은 컬렉션

† 정회원, 경상대학교 컴퓨터과학과/정보통신 연구센터
†† 정회원, 경남정보대학 컴퓨터 정보시스템 계열

을 선택 문제이다. 이는 메타 검색 시스템이 검색에 사용하고 있는 수많은 이질의 컬렉션 중에서 사용자의 질의를 만족시킬 수 있는 가장 좋은 컬렉션들을 자동으로 결정하는 방법이다. 두 번째는 질의어 자동 번역 문제이다. 메타 검색 시스템에서는 단일 인터페이스를 통해 질의어를 발생시킨다. 이때 발생된 질의어는 가장 적합한 컬렉션을 선택한 후에 자동적으로 질의된다. 그러나 검색에 참여한 이질의 컬렉션은 서로 다른 질의 문법을 가지고 있어서 메타 검색 시스템에서 생성된 질의어를 인식하지 못한다. 따라서 이들을 자동 번역하는 질의어 번역기가 필요하다. 마지막으로는 검색 문서의 통합 및 순위 매김하는 문제이다. 메타 검색 시스템은 입력된 질의어에 대하여 분산된 이질의 컬렉션으로부터 검색 결과를 수집한다. 그리고, 이들을 통합하고 문서에 대하여 순위 매김을 수행하여 사용자에게 단일 검색 결과를 제공한다. 이러한 메타 검색 시스템의 세 가지 연구 분야는 메타 검색 시스템의 검색 결과에 많은 영향을 미치며, 또한 검색을 수행할 때 상호 연동되어 동작한다. 만약, 검색에 참여하고 있는 컬렉션의 상세 정보가 많으면 많을수록 보다 양질의 검색 결과를 사용자에게 제공할 수 있다[5,7,8].

그러나 메타 검색 시스템에서 검색에 참여시키고 있는 서로 다른 이질의 컬렉션에서 질의에 적합한 검색 문서를 추출하고 하기란 매우 어렵다. 그 이유는 각 컬렉션에서 사용되는 문서 검색 알고리즘은 일반적으로 잘 알려져 있지 않다. 또한 특정 두 개의 컬렉션이 같은 문서 검색 알고리즘을 사용한다 하더라도 같은 질의로부터 나온 검색 결과에 대해서도 상대적으로 비교할 수 없다. 그 이유는 이들 컬렉션이 가지고 있는 전체 문서의 집합이 틀리기 때문에 같은 문서 검색 알고리즘을 사용한다 하더라도 같은 질의에 대하여 같은 결과가 나올 수 없기 때문이다.

본 논문에서는 메타 검색의 주요 연구분야 중 하나인 컬렉션 선택에 대한 모델을 제안한다. 이는 주어진 질의와 컬렉션사이의 관련성 분포정보를 이용한 컬렉션 선택 모델이다. 그리고 제안된 모델을 평가하기 위해 HoleInOne(*wHOLE INformation ONE-time*) 메타 검색기를 프로토 타입으로 구현하였으며, 실험을 통해 얻은 검색 결과는 검색 정확도 측면에서 15%정도 향상된 결과를 얻었다.

2. 관련 연구

컬렉션 선택은 메타 검색 시스템의 주요 연구 분야중 하나이다. 이는 메타 검색 시스템에서 질의가 주어졌을 때 분산된 컬렉션들 중에서 질의에 가장 적합한 컬렉션을 선택하는 것에 대한 문제이다. 따라서 이를 통해 메타 검색기에서는 어느 컬렉션으로부터 질의에 대해 문서를 검색할지 결정하게 되며, 이는 검색의 효율성에 많은 영향을 주는 한 요소다.

이 절에서는 먼저, 기존의 세 가지 컬렉션 선택 모델에 대해 살펴본다. 먼저 Voorhees[2,3]외 2명이 제안한 컬렉션 선택 모델, INQUERY[1,13] 메타 검색 시스템에서 사용한 컬렉션 선택 모델이고, 다음으로는 ProFusion[6] 메타 검색 시스템에서 사용한 컬렉션 선택 모델 등이다.

첫 번째로, Voorhees[2,3]외 2명이 제안한 컬렉션 선택 모델은 주어진 질의와 검색에 참여한 컬렉션과의 관련성을 유사도 값(similarity values)으로 평가하고, 이를 이용하여 컬렉션 선택을 결정하는 모델이다. 이때 컬렉션에 대한 유사도 값을 추정하는 방법으로는 문서의 관련성 분포(relevant document distribution)정보와 질의 클러스트링(Query Clustring) 정보를 이용하였다. 먼저, 문서의 관련성 분포 정보를 이용하는 방법은 먼저 질의들을 학습시켜 각 컬렉션에 대해 질의의 유사도 값을 평가하고, 이에 대한 정보를 저장한다. 만약 새로운 질의가 주어지면 질의와 유사한 k(임의의 갯수)개의 학습된 질의를 추출하여 이들이 가지고 있는 유사도 값들의 평균값을 새로운 질의에 대한 컬렉션의 유사도 값으로 추정하는 방법이다. 다음으로는 질의들의 클러스트링(Query Clustering) 정보를 이용하여 컬렉션을 선택하는 방법이다. 이는 앞의 방법과 동일하게 미리 질의들을 학습시켜 질의와 컬렉션사이의 유사도 값을 평가한다. 이렇게 학습된 질의들은 공통된 검색 문서의 빈도 수에 따라 질의들을 클러스트링 하며, 이들은 각각의 유사도 값들을 평균값으로 해당 컬렉션에 대한 유사도 값으로 추정하고 이를 중심 값(centroids values)이라 한다. 만약 새로운 질의가 입력되면 먼저 유사한 학습 질의를 찾고, 이 질의가 속해 있는 클러스트링의 중심 값을 새로운 질의에 대한 컬렉션의 유사도 값으로 평가하는 방법이다.

두 번째로는 Callan[1,13]외 3명이 제안한 모델로 *INQUERY* 메타 검색 시스템으로 실험을 하였다. 이는 *CORI net*(collection retrieval inference network) 검색 모델이라고도 하며, 문서(document), 컬렉션(collection)과 질의사이의 관련성을 *df*(document

frequency)와 *icf*(inverse collection frequency)를 기반으로 평가한다. 또한 질의와 컬렉션내의 문서 사이에 대한 관련성을 문서 네트워크 부분과 질의 네트워크 부분으로 분류하여 관련성 정보를 표현한 모델이다. *CORI net* 모델에서는 주어진 질의에 대하여 가장 적합한 컬렉션을 선택하기 위해 term과 *df*를 기반으로 컬렉션 선택 정보를 생성한다.

세 번째로는 *ProFusion*[6] 메타 검색 시스템에서 제안한 컬렉션 선택 모델이다. 이는 미국 캔자스 대학의 *Susan Gauch[6]*외 2명이 제안한 하였으며, 9개의 일반 검색 엔진을 대상으로 질의를 수행하고 이들로부터 검색 결과를 수집하여 통합 검색 결과를 사용자에게 URLs로 보여준다. *ProFusion* 메타 검색 시스템에서는 사용자의 질의에 대하여 9개의 컬렉션을 선택하는 방법으로는 최상의 3개 검색 엔진을 선택하는 방법, 가장 빠른 검색 결과를 보여주는 3개의 검색 엔진을 선택하는 방법, 9개의 검색 엔진 모두다 사용하는 방법, 사용자가 검색 엔진을 선택하여 사용하는 방법 등이 있으며, 본 논문에서는 최상의 검색 엔진을 판단하는 방법에 대해서만 다룬다. *ProFusion* 메타 검색 시스템은 최상의 3개 검색 엔진을 선택하기 위해 질의에 대하여 컬렉션을 평가하고 이를 기반으로 컬렉션을 선택할 수 있는 신뢰도(CF : Confidence Factor) 정보를 생성하고 이를 데이터베이스 정보로 구축하여 새로운 질의가 발생될 때 이를 사용한다.

먼저 최상의 검색 엔진을 선택하기 위해 뉴스 그룹에서 사용하는 도메인 네임으로부터 13개의 카테고리를 선정하여 이를 질의에 대한 분류로 사용하였다. 이에 대한 카테고리는 Science and Engineering, Computer Science, Travel, Medical and biotechnology, Business and Finance, Social and religion, Society, Law and goverment, Animals and Environment, History, Recreation and entertainment, Art, Music, Food 등이다. 그리고 이들 뉴스 그룹으로부터 4,000 개의 유일한 Term 후보들을 추출한 후에 이들 term이 카테고리내의 문서에 포함되어 있는 문서의 발생 빈도 수에 대한 정보를 지식 데이터베이스로 구축한다. 이러한 지식 데이터베이스 정보는 새로운 질의가 발생될 때 컬렉션 선택에 대한 신뢰도(cf) 값으로 사용된다. ProFusion에서는 컬렉션을 평가할 때에 두 가지 요소의 곱으로 한다. 첫 번째는 앞에서 언급한 컬렉션의 신뢰도(CF) 값이며, 다음은 검색된 문서의 우선 순위 정보(Ranking Factor)이다. 이는 검색된 문서가 가지는 우선 순위 값을 통해 질의와 관련된 문서의 우선 순위 값을 보상해준다. 이렇게 평가된 값을 기반으로 ProFusion 메타 검색기에서는 주어진 질의에 대해 최상의 3개 컬렉션을 선택한다.

3. 메타 데이터 기반 컬렉션 선택 모델

이 장에서는 질의와 컬렉션 사이의 관련성 정보를 사용하여 질의에 가장 적합한 컬렉션을 선택하는 새로운 모델을 제안한다. 이를 위해 3.1절에서는 본 논문에서 제시하는 컬렉션 선택 모델의 개괄 구조를 살펴보며, 3.2절에서는 컬렉션 선택 방법에 대해 살펴본다. 3.2절에서 제안된 컬렉션 선택 방법은 먼저, 질의에 대하여 컬렉션으로부터 검색된 문서의 요약 메타 데이터를 추출하고, 이를 사용하여 검색 문서를 평가한다. 그리고 평가된 검색 문서 정보를 기반으로 컬렉션에 대한 관련성 분포 정보를 추정하여 이를 기반으로 컬렉션을 선택한다.

3.1 컬렉션 선택의 개괄 구조

컬렉션 선택이란 메타 검색에서 주어진 질의에 대하여 가장 적합한 컬렉션을 선택하는 것을 말한다. 본 논문에서 제안하는 컬렉션 선택 모델은 그림 1과 같으며 이는 질의어 자동번역기, 검색문서 평가기, 컬렉션 평가기, 컬렉션 선택기 등 네 부분으로 구성되어 있다. 먼저 질의어 자동 번역기는 사용자로부터 주어진 검색 질의를 검색에 참여한 컬렉션들의 질의

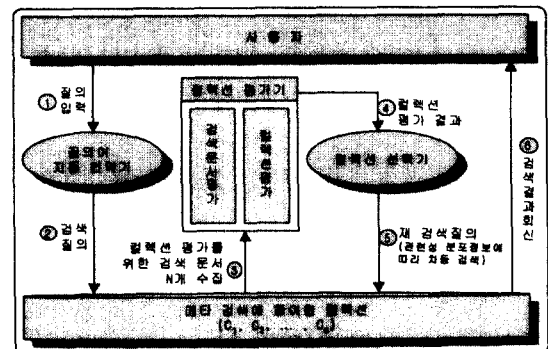



그림 1. 컬렉션 선택의 처리과정

문법에 맞게 자동 번역한 후에, 대신 질의하는 기능을 제공한다. 두 번째는 검색 문서 평가기로 검색에 참여한 컬렉션으로부터 회신된 검색 문서를 동일하게 검색된 문서의 우선 순위가 높은 크기 N만큼을 대상으로 본 논문에서 제안한 문서의 요약 메타 데이터를 추출하고, 제안한 문서 평가 방법에 따라 검색 문서를 평가한다. 이때 검색 문서를 평가하는 자세한 방법은 3.2.1절에서 소개한다. 세 번째는 컬렉션 평가기이다. 이는 질의에 대해 가장 적합한 컬렉션을 선택하기 위한 기준을 추정한다. 이를 위해 본 논문에서는 검색 문서의 재평가 값, 검색 문서의 위치 정보 평가 값, 검색 문서의 정확도 값 등으로 컬렉션을 평가하였다. 이에 대한 자세한 평가 방법은 3.2.2절에서 소개한다. 마지막으로 컬렉션 선택기이다. 이는 전 단계에서 수행한 컬렉션의 평가 값을 기반으로 질의에 가장 적합한 컬렉션을 선택하며, 또한 컬렉션의 상대적인 평가 값에 따라 해당 컬렉션으로부터 차등적으로 검색 문서를 수집한다.

다음의 그림 1은 본 논문에서 제안하는 컬렉션 선택 모델의 개괄적인 처리 과정이다.

이들의 처리 과정 순서를 화살표 위에 원 번호로 표시하였다. 먼저 사용자로부터 입력된 질의는 질의어 자동 번역기에 의해 검색에 참여한 컬렉션의 질의 문법으로 자동 번역되며, 질의를 통해 수집된 문서는 컬렉션 평가기에서 문서를 수집하여, (1) 검색문서 평가, (2) 컬렉션 평가를 통해 질의와 컬렉션 사이의 관련성 분포 정보를 평가한다. 평가된 관련성 분포정보는 컬렉션 선택기에서 질의에 대해 양질의 컬렉션으로 판단할 때 사용된다. 즉 사용자에게 검색 결과로써 검색 문서를 회신하기 위해 재 질의할 때, 컬렉션 평가 값의 상대적인 비율만큼만 검색문서를 수집함으로써 컬렉션을 차등적으로 평가하였다.

3.2 컬렉션 선택 방법

메타 검색 시스템은 사용자로부터 입력된 질의에 대해서 이질의 컬렉션으로부터 문서를 검색한 후에, 문서 검색 결과를 사용자에게 되돌려준다. 이때 메타 검색 시스템은 문서를 검색하기 전에 어느 컬렉션이 질의에 가장 적합한 문서를 가지고 있는지를 판단한 후 문서를 검색한다. 이러한 컬렉션에 대한 판단 과정을 컬렉션 선택이라고 한다. 만약 여러 컬렉션 중에서 질의에 가장 적합한 문서를 가지고 있는 컬렉션을 선택할 수 있다면 이는 매우 좋은 검색 결과를 사용자에게 제공해 줄 확률이 높다. 이를 위해 이 절에서는 양질의 컬렉션을 선택할 수 있는 검색 문서 평가 메타 데이터, 컬렉션 평가 메타 데이터와 이들 메타 데이터를 기반으로 하는 컬렉션 선택 방법을 제안한다. 먼저 3.2.1절에서는 검색 문서 평가를 위해 정의한 메타데이터와 검색 문서 평가 방법을 소개하고, 3.2.2절에서는 컬렉션을 평가하기 위한 메타데이터와 3.2.1절에서 제시된 검색 문서 평가를 통해 생성된 검색문서 평가 정보, 컬렉션 평가정보를 기반으로 컬렉션의 관련성 정도를 평가하는 방법을 소개한다.

3.2.1 검색 문서 평가

이 절에서는 검색된 문서를 평가하기 위해 정의한 메타데이터와 이를 바탕으로 검색 문서를 평가하는 방법을 소개한다. 이를 위해 본 논문에서는 3개의 문서 요약 메타데이터를 정의하였으며, 이를 기반으로 검색된 문서를 재평가하였다.

다음은 검색 문서를 재평가하기 위해 본 논문에서 정의한 문서 요약 메타 데이터이다.

■ Term Frequency : 이 메타 데이터는 검색된 문서에서 질의가 발생한 빈도 수에 대한 정보를 가지고 있다. 이는 검색문서가 질의와의 관련성을 평가할 때 사용되며, 이를 위해 <Tf> 메타 데이터 태그로 정의하였다.

■ Document Frequency : 이 메타 데이터는 컬렉션 평가를 위해 검색된 임의의 크기 N안의 문서 가운데 질의를 포함하고 있는 문서의 수에 대한 정보이다. 이는 검색 문서가 질의와의 관련성을 평가할 때 <Tf> 메타 데이터에 대한 보완 정보로 사용되며,이에 대한 메타 데이터 태크를 <Df>라 정의하였다.

■ Total Count : 이 메타 데이터는 컬렉션 평가를 위해 검색된 임의의 N에 대한 정보이다. 이는 검색문서가 질의와의 관련성을 평가할 때 <Tf> 메타 데이터에 대한 보완 정보로 <Df> 메타데이터와 함께 사용되며, 이에 대한 메타 데이터 태크를 <TCount>라 정의하였다.

다음으로는 앞에서 정의한 메타 데이터를 기반으로 검색 문서에 대한 관련성 분포 정보를 평가한다. 이때 처리되는 과정은 다음의 그림 2와 같다.

그림 2는 검색 문서를 재평가하는 과정이다. 먼저 검색 문서 평가기에서 검색된 문서의 메타 데이터

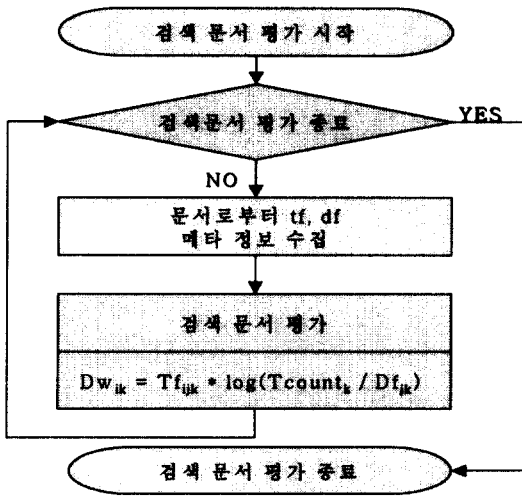



그림 2. 검색 문서의 평가 처리 과정

Tf, Df 및 TCount 메타데이터를 추출하며, 이러한 메타 데이터는 수식 1의 검색 문서 평가 방법에 따라 해당 검색 문서를 재평가한다. 이 평가 방법은 각 컬렉션에 서로 다른 문서 평가 방법을 사용하더라도 질의에 의해 검색된 문서는 서로 동등하다는 가정을 할 수 있다.

위에서 정의한 메타데이터를 기반으로 하는 검색 문서 평가 방법은 수식 1과 같다.

수식1은 전통적인 정보 검색에서 문서를 평가할 때 사용하는 수식이며, 본 논문에서는 3.2.1절에서 재정의한 문서 평가 메타데이터를 적용하여 검색 문서 재평가에 사용하였다. 수식 1에서 사용한 Tf_{ijk}는 검색된 문서 내에서 질의가 발생한 빈도 수에 대한 정보이다. 이는 하나의 문서 내에서 질의가 발생한 빈도 수에 따라 문서의 유사성이 높다는 측도로 사용된다. 그러나 일반적으로 해당 문서 내에서 검색 질의의 빈도 수가 너무 많으면 질의와는 관련이 없는 문서가 될 확률이 매우 높다. 따라서 이를 보완하

[수식 1]

$$DW_{ik} = \left(Tf_{ijk} * \log \frac{TCount_k}{Df_{jk}} \right)$$

- Df_{jk} : 컬렉션 k에서 term j 가 발생한 문서의 수
- Tf_{ijk} : 컬렉션 k의 문서 i 에서 term j 가 발생한 수
- $TCount_K$: 컬렉션 K에서 검색된 문서의 수

기 위해 $\log \frac{TCount_k}{Df_{jk}}$ 평가 항목을 사용하였다. 이는 검색 문서 내에서 발생하는 빈도 수가 너무 적은 경우와 너무 많은 경우에 질의와의 관련성 정도를 보완해주는 역할을 한다. 본 논문에서 사용한 수식 1의 문서 평가 방법은 전통적으로 tf와 idf를 이용하는 방법과 동일한 개념을 적용하였다. 이유는 이질의 컬렉션으로부터 문서를 검색할 때 이들 두 가지 메타데이터는 각 컬렉션에서 문서를 평가하는 알고리즘으로부터 독립적이다. 따라서 이들 두 가지 정보로 검색 문서를 재평가하면 이질의 컬렉션으로부터 검색된 문서라 할지라도 평가된 문서의 값을 동등하다고 할 수 있다. 이를 통해 얻어진 검색 문서의 평가값 DW_{ik}로 표현하였으며, 이를 컬렉션의 관련성 분포 정보를 평가할 때 하나의 요소로 사용하였다.

3.2.2 관련성 분포정보 평가

이 절에서는 검색 질의가 주어졌을 때, 가장 양질의 컬렉션이 어느 것인지를 판단하는 두 번째 과정이며, 그림 1의 컬렉션 평가기에서 이를 처리한다. 먼저 관련성 분포 정보란 검색 질의와 검색에 참여한 컬렉션사이의 관련성 정도를 말한다. 따라서 이 값이 크면 해당 컬렉션으로부터 검색 질의와 관련된 문서를 검색할 확률이 높다고 판단하며, 반대로 낮으면 검색 질의와 상관없는 문서를 검색할 확률이 높다고 판단하다. 이를 위해 여기에서는 3.2.1절의 검색 문서의 평가 결과를 기반으로 컬렉션에 대한 관련성 분포 정보를 평가하기 위한 3개의 메타데이터와 이에 대한 평가 방법을 제안한다.

다음은 컬렉션의 관련성 분포정보를 평가하기 위한 메타 데이터이다.

■ Document No : 이 메타 데이터는 메타 검색기에서 생성된 질의에 대하여 컬렉션으로부터 검색 문서를 수집할 때 검색 문서에 대한 우선 순위 정보 데이터이다. 이 메타데이터는 컬렉션을 평가하고, 문서에 대한 순위 매김을 할 때 각각 사용되며, 이를 위해 <DNo> 메타 데이터 태그를 정의하였다.

■ Document URL : 이 메타 데이터는 컬렉션에서 검색된 문서의 인터넷 주소 데이터를 가지고 있다. 본 논문에서는 이질의 컬렉션으로부터 문서를 검색한 후에 검색 문서의 URL을 분석하여 검색 문서의 중복성 및 빈 URL 등을 판단할 때 사용한다. 만약

검색된 문서 가운데 중복된 URL이 발생되면 동일한 문서를 검색한 것으로 간주하여 중복된 문서를 삭제할 때 사용하고, 빈 URL인 경우에는 관련없는 문서로 간주할 때 사용하였다. 이를 위해 <DUrl> 메타 데이터 태그를 정의하였다.

■ Information Types : 이 메타 데이터는 검색에 참여한 컬렉션의 종류에 대한 메타 데이터이다. 이는 컬렉션에 대하여 검색 문서의 관련성을 평가할 때 평가 요소로 사용된다. 이를 위해 <IT> 메타 데이터 태그를 정의하였다.

컬렉션에 대한 관련성 분포 정보 평가는 3.2.1절의 검색 문서 평가 결과와 앞에서 정의한 메타 데이터를 기반으로 평가한다. 이때 평가되어지는 개괄적인 처리과정은 다음의 그림 3과 같다.

그림 3에서 컬렉션에 대한 관련성 분포 정보 평가는 크게 세 가지 요소를 기반으로 평가를 한다. 먼저 검색 문서의 평가 값에 대한 합이다. 이는 3.2.1절에서 검색 문서 평가 방법을 소개하였으며, 검색 문서 평가 값은 질의와의 관련성 정도를 나타낸다고 할 수 있다. 따라서 평가된 검색 문서의 결과 값을 합하여 컬렉션을 평가할 때 사용하였다.

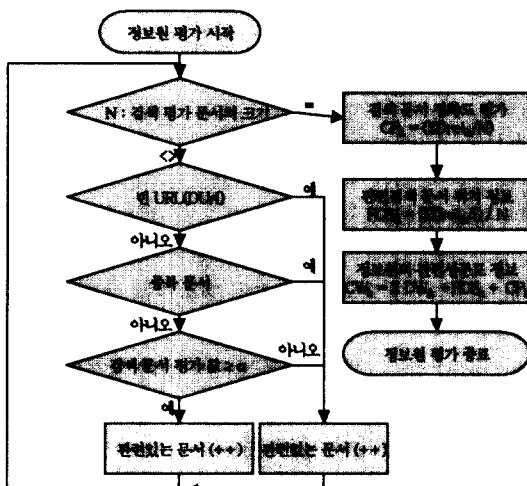
두 번째는 검색 문서의 정확도이다. 이는 컬렉션을 평가할 때 평가하고자하는 검색 문서 가운데 유일하게 관련 문서로 판단된 비율을 정확도라 한다. 이때 검색 문서 가운데서 중복된 문서, 빈 URL, 검색문서 평가 값이 α보다 작은 경우 등은 관련없는 문서로 간주하였다. 먼저 중복된 문서의 판별은 검색된 문서가 가지는 URL을 서로 비교하여 동일한 URL일 경우에는 중복된 문서로 간주하였으며, 빈 URL의 판단은 검색된 문서가 가지는 URL을 사용하여 메타 검색기가 문서 요청 메시지를 전달하여 되돌아오는 신호를 보고 판단하였다. 본 논문에서는 문서 요청에 대한 HTTP 회신 코드가 "4xx"일 경우에는 현재 사용할 수 없는 URL로 이 정보를 빈 URL 구분에 사용하였다.

그림 3. 컬렉션의 관련성 평가 처리 모델

이에 대한 평가 과정은 다음의 수식 2로 정의하였다.

수식 2에서 $Drel_{ik}$ 은 검색 문서의 평가 값을 통해 검색된 문서와 질의 사이의 관련성에 대한 정보를 가지는 메타 데이터이다. 이때 관련성 메타데이터 값이 1 일 경우는 관련 문서로, 0일 경우는 관련없는 문서로 판단하였으며, 이를 통해 해당 컬렉션에서 관련 문서의 수를 합한 후에 평가 대상 전체 문서의 수 N으로 나누어 컬렉션에 대한 정확도 CP_k 값으로 계산하였다.

마지막으로 관련 문서의 위치 정보에 대한 평가 값이다. 이는 컬렉션을 평가할 때 관련있는 문서의 위치에 따라 컬렉션의 관련성 정도를 보상해 주기 위해 사용한다. 다음의 수식 3은 컬렉션에 대한 위치 정보 평가식이다.

수식 3에서 $Drel_{ik}$은 질의와 검색 문서 사이의 관련성 값을 가지며, 이를 검색 문서가 가지고 있는 위

[수식 2]

$$CP_k = \frac{\sum_{i=1}^{N} Drel_{ik}}{N}$$

- $Drel_{ik}$: 컬렉션 k에서 문서 i에 대한 관련성 검사 값
- CP_k : 컬렉션 k의 정확도 값

[수식 3]

$$POS_k = \frac{\sum_{i=1}^{N} \frac{Drel_{ik}}{i}}{N}$$

- 컬렉션 k에서
- $Drel_{ik}$: 컬렉션 k 에서 문서 i에 대한 관련성 검사 값
- POS_k : 컬렉션 k에서 관련문서에 대한 위치 정보 값

치 정보 i 를 사용하여 $\sum_{i=1}^{N}\frac{Drel_{ik}}{i}$ 를 계산하였다. 이는 컬렉션을 평가하기 위해 정의한 모집단에서 관련문서의 위치 정보에 대한 가중치이다. 이를 모집단의 크기 N으로 나누어 최종적으로 컬렉션 평가에서 관련 위치 정보의 가중치 값으로 사용하였다.

따라서 검색에 참여한 컬렉션이 주어진 질의에 대하여 얼마나 좋은 문서 집단인지를 나타내는 척도를 관련성 분포 정보라 하였다. 즉 질의에 대해서 양질의 문서를 많이 가질 경우에는 이 값이 커지며, 반대로 질의와 관련이 적을수록 관련성 분포정보도 작아진다. 이를 위해 수식 1, 2, 3에서 계산한 값을 합하여 해당 컬렉션에 대한 평가 값으로 정의하였다. 다음의 수식 4는 컬렉션에 대한 관련성 분포 정보를 계산하는 식이다.

수식 4에서 사용한 구성 요소는 먼저 수식 1에서 계산한 문서에 대한 Term Weight 가중치(DW_{ik})와 수식 2에서 평가한 관련 문서의 정확도(CP_k) 그리고 수식 3에서 계산한 관련 문서에 대한 위치 정보 값 (POS_k) 등으로 컬렉션에 대한 관련성 분포 정보를 평가하였다.

다음으로는 질의에 대하여 평가된 컬렉션의 개별적인 관련성 분포 정보 값을 검색에 참여한 모든 컬렉션에 대하여 상대적인 관련성 분포 정보 비율을 계산한다. 이는 검색에 참여한 컬렉션 사이의 상대적인 평가 자료가 된다.

수식 5에서 사용한 CW_k는 수식 4에서 평가된 컬렉션의 가중치 값이다. 이 CW_k 값을 사용하여 컬렉션에 대한 상대적 관련성 분포정보를 계산하였다. 즉 질의에 대하여 평가된 개별 컬렉션의 가중치 값에 검색에 참여한 모든 컬렉션에 대한 가중치 값을 합산한 값 $\sum_{k=1}^{m} CW_k$ 를 나누어 개별 컬렉션이 차지하는 상대적인 비율을 얻을 수 있다. 이를 통해 질의에 대하여 해당 컬렉션이 차지하는 상대적인 비율을 계산하였다.

[수식 4]

$$CW_k = \left[\sum_{i=1}^{N} DW_{ik} + POS_k + CP_k\right]$$

- DW_{ik} : 컬렉션 k 에서 문서 i 에 대한 term 가중치 값
- POS_k : 컬렉션 k 에 대한 위치 정보 값
- CP_k : 컬렉션 k 에 대한 정확도

[수식 5]

$$CW_{Rate}k = \frac{CW_k}{\sum_{k=1}^{m} CW_k}$$

- $\sum_{k=1}^{m} CW_k$: 각 컬렉션 가중치 값의 합계
- CW_k : 각 컬렉션 가중치 값
- m : 검색에 사용된 컬렉션의 수

본 논문에서는 평가된 컬렉션의 관련성 분포정보를 사용하여 각 컬렉션으로부터 검색 문서의 수를 상대적 비율만큼 수집하였다. 또한 이질의 정보원으로부터 검색된 문서들을 하나의 검색 결과 집합으로 통합할 때에는 관련성 분포정보의 상대적 비율을 기반으로 동일한 비율 내에 포함된 검색문서는 문서의 순위 매김 방법이 동등하다는 것으로 가정하였다.

다음의 그림 4는 수식 5에서 계산된 관련성 분포정보의 상대적 비율을 기반으로 3개의 컬렉션으로부터 검색 문서를 수집하는 예제이다.

예를 들어 그림 4와 같이 A, B, C 3개 컬렉션이 검색에 참여하였으며, 이들의 상대적 관련성 분포정보가 60%, 30%, 10%로 평가되었다고 가정한다. 이때 각 컬렉션으로부터 검색할 문서의 수는 A 컬렉션으로부터 6개, B 컬렉션으로부터 3개, C 컬렉션으로부터 1개 등의 비율로 문서를 검색하였으며, 검색된 문서들은 점선 화살표로 표시된 부분 내에서 문서들의 우선 순위를 정하여 이를 내림차순으로 통합하였다.

4. 비교 분석

이 장에서는 기존의 컬렉션 선택 모델과 본 논문

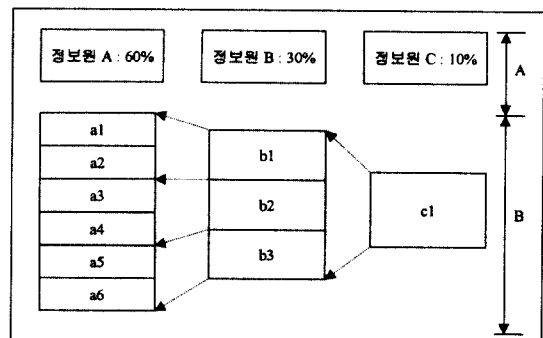



그림 4. 관련성 분포정보를 이용한 문서수집 방법

에서 제안한 컬렉션 선택 모델에서 사용된 메타 데이터와 이들의 특징들을 비교 분석해본다. 아래의 표 1은 컬렉션 선택에서 사용된 메타 정보들을 서로 비교해 보고, 그 특징들에 대해 간략히 기술하였다.

현재까지 알려진 대부분의 기존 연구들은 주어진 질의에 가장 적합한 컬렉션을 선택하기 위해 자신의 검색 정보를 새롭게 생성하고, 이들을 평가 기준으로 질의가 발생될 때 사용하였다. 이는 인터넷상에 존재하는 수많은 문서들에 대해 검색 정보를 생성해야 한다는 단점이 있으며, 또한 생성된 검색 정보들을 동적으로 변하는 인터넷 환경에서 일관성 유지를 위해 사용되는 비용도 매우 크다. 그래서 본 논문에서는 이러한 검색 정보를 생성하지 않으면서, 검색의 효율을 보장할 수 있는 메타 정보를 사용하여 컬렉션을 선택하는 모델을 제안하였다.

본 논문에서 제안하는 모델의 가장 큰 특징은 (1) 검색 정보를 생성하지 않고, (2) 메타 정보만으로 컬렉션을 선택할 수 있는 장점이 있다.

5. 실험결과

이 장에서는 본 논문에서 제안한 컬렉션 선택 모델에 대해 성능을 평가하기 위해 구현한 *HoleInOne* 메타 검색 시스템과 현재 일반적으로 많이 사용되고 있는 5개의 일반 검색 엔진들과 3개의 메타 검색 엔진과의 실험 결과를 소개한다.

먼저, 검색 평가를 위해 사용한 질의와 평가 항목들은 다음과 같다. 검색 질의로는 ProFusion 메타 검색에서 평가를 위해 사용한 뉴스 그룹에서의 13개의 주제어를 검색 질의로 사용하였다. 이들 질의는 Science and engineering, Computer Science, Travel, Medical and Biotechnology, Business and Finance, Social and Religion, Society Law and Government, Animals and Environment, History, Recreation and Entertainment, Art, Music, Food 등이다. 이러한 질의는 일반적으로 인터넷상의 자료들을 주제별로 분류할 때 자주 사용되는 기준이다. 즉 이들 질의는 특정한 주제에 종속되지 않아 보편성을 가지고 있으며, 질의로 사용되어도 임의의 검색 시스템에 종속되는 결과는 발생하지 않는다고 가정할 수 있다. 다음은 검색 결과 평가를 위한 항목들로써 (1) 관련 문서의 수, (2)비 관련 문서의 수, (3)빈 URL, (4)중복 관련 문서 문서의 수, (5)유일 관련 문서의 수, (6)검색 문서의 정확도 등을 사용하였다. 본 논문

표 1. 컬렉션 선택에서 사용한 메타정보 비교

<table>
<tr><th colspan="2" rowspan="2">구 분</th><th colspan="2">컬렉션 평가요소</th><th rowspan="2">특징</th></tr>
<tr><th>평가정보</th><th>검색 데이터베이스</th></tr>
<tr><td rowspan="2">Voorhees 외 2명 제시모델</td><td>관련성분포 정보</td><td>Df</td><td>1) 질의의 관련성 분포정보 DB</td><td rowspan="2">■ 장점
1) 빠른 컬렉션선택
■ 단점
1) 질의를 학습시키기가 힘들다.
2) 학습 질의로 검색 정보의 관리가 힘들다.</td></tr>
<tr><td>클러스트링</td><td>Df</td><td>1) 질의의 클러스트링 검색 정보 DB</td></tr>
<tr><td colspan="2">INQUERY</td><td>Df
Icf</td><td>1) CORI net의 검색 DB 정보</td><td>1) 검색 인덱스 정보 생성 및 유지가 힘들다.</td></tr>
<tr><td colspan="2">ProFusion</td><td>Df</td><td>1) 지식DB 검색정보</td><td>■ 장점
1) 빠른 컬렉션선택
■ 단점
1) 지식 DB를 매일 갱신한다.
2) 지식 DB의 관리 비용이 크다.</td></tr>
<tr><td colspan="2">본 논문의 제안 모델</td><td>Tf,
Df,
N</td><td>1) 사용하지 않음</td><td>■ 장점
1) 검색 인덱스 정보 생성비용이 없다.
2) 동적 환경에 적용하기 쉽다.
■ 단점
1) 컬렉션 평가를 컬렉션을 선택할 때 처리하여 검색에 대한 응답 시간이 늘어난다.</td></tr>
</table>

에서는 (1)번과 (5)번 항목으로 컬렉션에 대한 검색 결과의 정확성을 평가하였다.

아래의 표 2는 컬렉션으로부터 검색된 문서 중에서 상위 30개의 문서를 대상으로 (1) 관련된 문서, (2) 비 관련된 문서, (3)빈 URL, (4) 중복 검색된 문서, (5) 유일하게 관련된 문서, (6) 정확도 등의 6가지 항목으로 분류하여 컬렉션 선택의 효율성을 간접적으로 비교하였다.

본 실험에서는 제안한 관련성 분포 정보를 기반으로 검색 결과와 기존의 단일 검색엔진 결과와의 비교에서 컬렉션의 선택 면에서 약 15%의 성능 향상을 확인하였다.

6. 결 론

인터넷상의 수많은 컬렉션들은 독자적인 정보관리 모델을 가지고 있다. 이는 인터넷의 폭넓은 보급으로 인터넷의 바다에서 정보를 찾고자하는 사용자들에게 다양한 컬렉션들의 검색 방법 사용에 대한 추가적인 부담되고 있다. 따라서 인터넷에서 정보를 검색하고자할 때 쉽고 편리하면서도 찾고자하는 정보를 정확하게 수집할 수 있는 검색 방법이 요구되고 있다. 본 논문에서는 통합 검색 시스템을 사용하여 인터넷상의 다양한 정보들을 효율적으로 검색할 수 있는 모델을 제안하고, 이 제안된 모델의 검색 성능을 평가하기 위해 *HoleInOne* 통합 검색 시스템을 설계 및 구현하였다.

향후 연구 과제로는 컬렉션에 대한 양질의 정보를 얻기 위해서 질의에 적합한 컬렉션을 선택할 수 있도록 표준화된 메타 데이터 개발, 컬렉션 평가를 위한 컬렉션 평가 메타데이터 설계 및 개발, 컬렉션에 대한 정보 수집 방법과 융합 클러스터링 기법의 개발 등에 대한연구가 필요하다. 또한 질의 처리 기능의 확장이 필요하다. 즉, 불리언 모델에 바탕을 둔 질의 처리 기능과 순위 매김(ranking) 모델에 바탕을 둔 질의 처리 기능 등의 연구이다. 이러한 정보는 본 논문에서 제시된 알고리즘의 성능을 크게 개선시킬 수 있다.

표 2. 컬렉션에서 검색된 문서 중 상위 30개에 대한 분석

Search Engines	Number of relevant	Number of irrelevant	broken links	duplicate relevant document	nuber of unique relevant document	precision number unique
Single Search Engines						
WebCrawler	230	147	13	8	222	0.57
Yahoo	245	138	7	19	226	0.58
InfoSeek	300	80	10	3	297	0.76
Excite	270	110	10	4	266	0.68
AltaVista	245	103	42	4	241	0.62
Meta Search Engines						
SavvySearch	211	160	19	15	196	0.50
ProFusion	234	133	23	11	223	0.57
미스다찾니	226	146	18	6	220	0.56
HoleInOne	335	25	0	13	322	0.83

참 고 문 헌

[1] J. P. Callan, Z. Lu, and W. B. Croft, "Searching Distributed Collections with Inference Networks," In Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA. pp. 21-28. 1995.

[2] E. M. Voorhees, N. K. Gupat, and B. Johnson-Laird., "The Collection Fusion Problem," In D. K. Harman, editor, The Third Text REtrieval Conference (TREC-3), Gaithersburg, MD, pp. National Institue of Standards and Technology, Special Publication 500-225., 1994.

[3] E. M. Voorhees, N. Gupta, and B. Johnson-Laird., "Learning Collection Fusion Strategies," SIGIR '95, p172-179, 1995.

[4] C. L. Viles and J. C. French, "Dissemination of Collection Wide Information in a Distributed Information Retrieval System," ACM SIGIR 95, 1995.

[5] A. Moffat and J. Zobel, "Information Retrieval Systems for Large Document Collections," TREC-3, pp85-94., 1994.

[6] S. Gauch, G. Wang, and M. Gomez, "ProFusion : Intelligent Fusion from Multiple, Distributed Search Engines, WebNet '96", The First World Conference of the Web Society, San Francisco, October 1996.

[7] C. Baumgarten, "Probablistic Information Retrieval in a Distributed Heterogeneous Environment." PhD Thesis, Dresden Univ. of Techn., Accepted, 1999.

[8] C. Baumgarten, "A Probabilistic Solution to the Selection and Fusion Problem in Distributed Information Retrieval," SIGIR '99, 1999.
[9] J. C. French, A. L. Powell, J. Callan, C. L. Viles, T. Emmitt, K. J. Prey and Y. Mou, "Comparing the Performance of Database Selection Algorithms," SIGIR 99, 1999.
[10] N. Fuhr, "Resource Discovery in Distributed Digital Libraries," SIGIR 99, 1994.
[11] L. Gravano, C. K. Chang, H. Gracia-Molina, and A. Paepcke. "STARTS: Stanford protocol proposal for Internet retrieval and search," Technical Report SIDL-WP-1996-0043, Stanford University, August, 1996.
[12] J. Xu and J. P. Callan. "Effective Retrieval with Distributed Collections." In Proceedings of the Twenty-first Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, August 1998.
[13] L. Gravano, H. Garcia-Molina and A. Tomasic, "The effectiveness of GlOSS for the text database discovery problem," In Proceedings of the 1994 ACM SIGMOD Conference, May 1994.
[14] 금기문, 남세진, 신동욱, 김태균, "문서 클러스터링 정보를 이용한 컬렉션 융합",한국정보과학회 추계학술 논문발표집. pp.147-149., 1998.
[15] 김현주, 김상준, 배종민 "관련성 분포정보를 이용한 컬렉션 융합", 한국 정보처리학회 '99봄 학술발표 논문집(하), pp.907-910., 1999.
[16] 김연곤, 엄채임, 변정용, "빈 연결을 제거하는 메타 검색 엔진의 구현," 한국멀티미디어학회 추계학술발표회, pp.359-364., 1998.
[17] 신봉기, 김영환, "인터넷 정보검색 서비스 동향," 한국정보과학회지 제16권 제8호, pp.16-20., 1998.

배 종 민

1980년 서울대학교 사범대학 수학과(이학사)
1983년 서울대학교 계산통계학과(이학석사)
1995년 서울대학교 계산통계학과(이학박사)
1982년~1984년 한국전자통신연구원 연구원
1984년~현재 경상대학교 컴퓨터과학과 교수
관심분야 : 병렬 프로그래밍 언어, 디지털 도서관, 정보검색

김 현 주

1988년 경상대학교 전산통계학과(이학사)
1990년 숭실대학교 전자계산학과(공학석사)
2000년 경상대학교 전자계산학과(공학박사)
1994년~1997년 제일정밀공업(주)연구원
20000년~현재 경남정보대학 컴퓨터정보계열 전임강사
관심분야 : 정보검색, 디지털 도서관, 웹 프로그래밍